UNIVERSITY OF WOLVERHAMPTON

HERALD COLLEGE KATHMANDU

Data Visualization

Project Report

HR & Workforce Analysis

Group : 3

Group Name : Elixir Golem

Members : Shashank Pandey

: Puskar Koirala

: Sambhu Kamti

Submitted on : 09-15-2025

Table of Contents

Abstract

Employee attrition represents one of the most pressing challenges faced by organizations, as it results in increased recruitment costs, the loss of valuable skills, and the disruption of operations. This project applies a data-driven approach to understanding attrition using the IBM HR Analytics Employee Attrition dataset. Following the workflow of a complete analytics pipeline, the dataset was cleaned, explored, and statistically analyzed to uncover significant factors influencing employee turnover. Hypothesis testing confirmed that attrition is significantly associated with job role, department, overtime status, monthly income, and years at the company. Predictive models including logistic regression, decision tree, random forest, support vector machine (SVM), and gradient boosting were evaluated, with logistic regression and random forest performing best. Feature importance analysis revealed that overtime, monthly income, and job role are the strongest predictors of attrition. An interactive dashboard was also developed in Plotly Dash, enabling stakeholders to explore KPIs, attrition patterns, and predictive model results in real time. The findings suggest that organizations can reduce attrition by redesigning overtime policies, revising compensation structures, and implementing role-specific retention strategies.

Introduction

Human capital is central to organizational competitiveness, yet high turnover rates remain a costly and disruptive issue. Employee attrition not only leads to direct expenses related to recruitment and training but also causes intangible costs such as decreased morale, lowered productivity, and knowledge loss. As organizations face dynamic markets and tight labor conditions, predicting and preventing attrition has become a priority for strategic human resource (HR) management.

The purpose of this project is to identify factors influencing attrition and to construct predictive models that enable HR managers to take proactive actions. By applying statistical analysis and machine learning to the IBM HR Analytics Employee Attrition dataset, this study seeks to answer key business questions regarding the relationship between attrition and factors such as job role, department, overtime, monthly income, and tenure. The project integrates exploratory data analysis (EDA), hypothesis testing, predictive modeling, and dashboard development to provide a comprehensive HR analytics solution.

Business Problem Definition

Employee attrition disrupts organizational continuity. The business problem addressed in this project is how to identify the drivers of attrition and predict which employees are most at risk.

Based on HR theories and prior research, the following business questions and hypotheses were formulated:

1. Is attrition rate significantly different across job roles?
   - H_0: Attrition does not differ across job roles.
   - H_1: Attrition differs significantly across job roles.
2. Is attrition rate significantly different across departments?
   - H_0: Attrition does not differ between departments.
   - H_1: Attrition differs significantly between departments.
3. Do employees who work overtime have higher attrition compared to those who don't?
   - H_0: Overtime is unrelated to attrition.
   - H_1: Overtime is significantly related to attrition.
4. Does monthly income differ significantly between employees who left and those who stayed?
   - H_0: No income difference exists between leavers and stayers.
   - H_1: Monthly income differs significantly between leavers and stayers.
5. Is there a significant relationship between years at the company and attrition?
   - H_0: No relationship exists between tenure and attrition.
   - H_1: Tenure is significantly related to attrition.

These hypotheses were tested using chi-square tests of independence, t-tests, and ANOVA where appropriate.

Dataset Description

The dataset used is the IBM HR Analytics Employee Attrition dataset from Kaggle. It contains 1,470 employees with 35 variables, including both categorical and numerical attributes. Examples include:

- Categorical: JobRole, Department, OverTime, Attrition.
- Numerical: Age, MonthlyIncome, YearsAtCompany, JobSatisfaction.

This dataset is highly suitable as it meets project requirements: it has a clear business domain (HR analytics), contains both categorical and numerical variables, and has over 500 records for robust statistical testing and modeling.

Data Preparation

Data preparation involved the following steps:

- Missing Values: No significant missing values were found. Where present, categorical values were imputed using mode and numerical values using median.
- Outliers: Outliers in income and years at company were retained as they reflect genuine business conditions.
- Encoding: Categorical variables were converted into numerical form using OneHotEncoding.
- Scaling: Numerical variables such as income were standardized for use in models like logistic regression and SVM.

The resulting dataset was clean, consistent, and ready for analysis.

Exploratory Data Analysis (EDA)

EDA provided an initial overview of attrition patterns:

- Attrition Rate: 16% of employees had left the company.
- Departmental Differences: Sales had an attrition rate of 20.6%, compared to 14.2% in R&D.
- Job Role: Laboratory Technicians and Sales Executives exhibited the highest attrition rates.
- Overtime: Employees working overtime were disproportionately more likely to leave.
- Income and Tenure: Employees who left typically earned lower salaries and had shorter tenures.

Visualizations included count plots, boxplots, histograms, and correlation heatmaps. These revealed clear patterns that guided hypothesis testing and modeling.

Hypothesis Testing

To statistically validate business assumptions:

- Department × Attrition: $\chi^2(2, N=1470) = 10.80$, $p = 0.0045$. Attrition significantly differs between departments.
- JobRole × Attrition: $\chi^2(8, N=1470) = 86.19$, $p < 0.001$. Strong evidence that attrition varies by job role.
- OverTime × Attrition: Highly significant association ($p < 0.001$). Overtime workers face higher attrition.
- Monthly Income: Independent samples t-test confirmed employees who left had significantly lower income ($p < 0.01$).
- Years at Company: ANOVA revealed tenure differences between leavers and stayers were statistically significant.

Effect sizes (e.g., Cramér’s V) were small to moderate, but practically meaningful for HR decision-making.

Predictive Modeling

Five machine learning models were trained and evaluated:

1. Logistic Regression
    - Accuracy: 76%
    - Recall: 68%
    - ROC-AUC: 0.81
    - Advantage: interpretable coefficients.
2. Random Forest (tuned)
    - Accuracy: 85%
    - Recall: 65%
    - F1 Score: 0.52
    - Advantage: high accuracy and ability to capture nonlinear patterns.
3. Decision Tree
    - Simpler, but prone to overfitting; inconsistent results.
4. Support Vector Machine (SVM)
    - Moderate performance, required heavy tuning.
5. Gradient Boosting
    - Performed moderately well but not superior to Logistic Regression or Random Forest.

Best models: Logistic Regression (interpretability) and Random Forest (performance).

Feature Importance (Random Forest):

- Top predictors: OverTime, MonthlyIncome, JobRole, Age, JobSatisfaction.

Dashboard Development

An interactive dashboard was built using Plotly Dash. Key features:

- **KPIs**: Attrition rate, income distributions, overtime status, attrition by department/job role.
- **Interactive Filters**: Users can drill down by department, job role, and tenure.
- **Model Results**: Displays logistic regression and random forest predictions, with probability scores for attrition risk.

This dashboard transforms static findings into a practical HR tool for data-driven decision-making.

Conclusion

This project demonstrated a complete HR analytics pipeline: from business problem definition to EDA, hypothesis testing, predictive modeling, and dashboard development.

Key conclusions:

- Attrition significantly differs across job roles, departments, and overtime status.
- Low income and short tenure are strong predictors of attrition.
- Logistic regression and random forest provide robust predictive capability.
- Overtime, income, and job role are the most influential features.

Recommendation: HR departments should reduce overtime pressures, ensure competitive salaries, and focus retention efforts on high-attrition roles and departments. Predictive models can be integrated into HR systems to flag at-risk employees early, enabling proactive interventions.

By leveraging data analytics, organizations can transform attrition from a reactive problem into a predictable and manageable process, improving employee satisfaction and ensuring organizational stability.